

‘Data Mining’

Data Mining Project with


Master 1 MLDM

Saint-Étienne, France

Fabrice Muhlenbach

Laboratoire Hubert Curien, UMR CNRS 5516
Université Jean Monnet de Saint-Étienne
18 rue du Professeur Benoît Luras
42000 SAINT-ÉTIENNE, FRANCE
<http://perso.univ-st-etienne.fr/muhlfabr/>

1 Objectives

The data mining project is not about finding a “good” dataset. The objective is (1) to search a real life and/or everyday life problem and (2) to find a sufficiently large dataset for being able to apply data mining techniques with  to find some answers to this problem.

An important part of this project is to find “amazing knowledges” (interesting, unexpected, or valuable structures) that are embedded in a large dataset.

2 Report

The data mining project report must be a 6-page document (PDF version) including an analysis describing the following six points: (1) Problem Understanding, (2) Data Understanding, (3) Data Preparation, (4) Modeling, (5) Evaluation, and (6) Deployment.


3 Datasets

With the appearance of the “Open Data”, it is now possible to find data on multiple subjects. For example, you can find some datasets here:



- **Kaggle** website: a platform for predictive modeling and analytics competitions on which companies and researchers post their data and statisticians and data miners from all over the world compete to produce the best models. For example :
 - Football Events: More than 900,000 events from 9,074 football games across Europe. You can use this dataset for example to create predictive models for football games in order to bet on football outcomes
 - Climate Change: Earth Surface Temperature Data: Exploring global temperatures since 1750. You can use this dataset to see the temperature evolution in a given geographic area and create a predictive model of the temperature or the sea level in the next decades

- Official open data portals for some countries or institutions, for example:
 - French public open data platform
 - Open data France: association of local authorities
 - Open Data in the U.K.
 - U.S. Government’s open data
 - UNICEF (United Nations Children’s Fund): situation of women and children worldwide
 - Some cities: Chicago, Paris, Lyon...
- Less interesting: UC Irvine Machine Learning Repository: datasets often too small to apply interesting data mining techniques, the problems have already been studied by many data scientists, therefore it is difficult to find some surprising elements.

4 Storage

Your  code as well as the dataset must be hosted on a GitHub repository.

Note that you can easily work with GitHub within RStudio (see for example this tutorial).

Moreover you can create your own  package. An  package stored on GitHub can be automatically installed with `devtools` or `githubinstall` packages (see for example this tutorial).

5 Date

You must send your report before Friday, March the 23th 2018, 11:59 PM (UTC+1, Paris or Saint-Etienne local time) by electronic mail to your teacher: fabrice.muhlenbach@univ-st-etienne.fr (note that the enclosed document must have less than 5 MB).

You have to send a copy of this report before the same deadline on the “Assignments” folder (entitled “Data Mining Project”) in the class **DM_LAB_M1** on *Claroline* platform.